Linear models for classification

Lecture 6

$\vec{x} \rightarrow C_k \,, \quad k = 1, \ldots, K$

input, D dims ... discrete classes

Input space divided into decision regions by decision surfaces.

For ex., $\vec{t} = \underbrace{(0, 1, 0, 0, 0)}$

target variables ... K=5 classes, target variable indicates class 2

Classification approaches:

1. Discriminant function : directly assigns $\vec{x}$ to a class, e.g. for 2 classes $y(\vec{x}) \geq 0 \iff C_1$, $\quad y(\vec{x}) < 0 \iff C_2$.

2. Probabilistic approach:

use $p(C_k|\vec{x}) = \underbrace{\dfrac{p(\vec{x}|C_k)\,p(C_k)}{p(\vec{x})}}_{\text{Bayesian framework}}$

Previously, we focused on

$y(\vec{x}) = \vec{w}^T \vec{\varphi}(\vec{x}) + w_0$. ← explicit bias prm

Now we will consider $y(\vec{x}) = f(\vec{w}^T \vec{x} + w_0)$

or, more generally, $y(\vec{x}) = f(\vec{w}^T \cdot \vec{\varphi}(\vec{x}) + w_0)$

non-linear activation f'n

Decision surfaces are given by

$\vec{w}^T \vec{x} + w_0 = \text{const}$, linear f's of $\vec{x}$

Discriminant functions

(1.) Two classes

Consider $y(\vec{x}) = \vec{w}^T \cdot \vec{x} + w_0$ [linear discriminant]

$$\begin{cases} y(\vec{x}) \geq 0 \Rightarrow C_1 \\ y(\vec{x}) < 0 \Rightarrow C_2 \end{cases}$$

$y(\vec{x}) = 0 \Leftarrow$ decision boundary (DB)

Consider $\vec{x}_A, \vec{x}_B \in DB$

$$y(\vec{x}_A) = y(\vec{x}_B) = 0 \Rightarrow \vec{w}^T (\underbrace{\vec{x}_A - \vec{x}_B}) = 0$$

lies on $(D-1)$ dim'l DB

So, $\vec{w} \perp DB \Rightarrow \vec{n} = \frac{\vec{w}}{\|\vec{w}\|}$ is a unit vector $\perp DB$

Similarly, if $\vec{x} \in DB$,

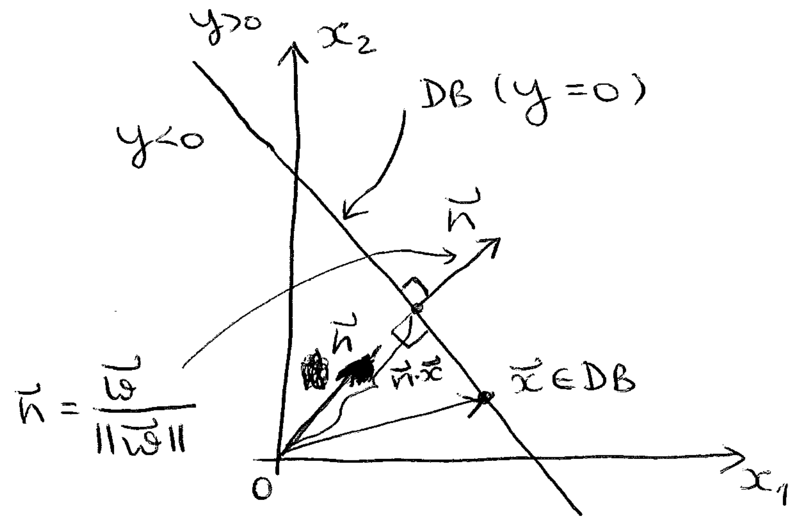
$$y(\vec{x}) = \vec{w}^T \cdot \vec{x} + w_0 = 0 , \text{ or}$$

$$\underbrace{\frac{\vec{w}^T \cdot \vec{x}}{\|\vec{w}\|}} = - \frac{w_0}{\|\vec{w}\|}$$

" $\vec{n} \cdot \vec{x}$ = normal distance from the origin to DB

So, w_0 determines the location of DB

D=2



Moreover, for any $\vec{x}$

$$\vec{x} = \vec{x}_{\parallel DB} + \underbrace{r\vec{n}}_{\perp DB}$$

Then

$$\underbrace{\vec{w}^T\vec{x} + w_0}_{y(\vec{x})} = \vec{w}^T\vec{x}_{\parallel DB} + r\,\frac{\overbrace{\vec{w}^T\cdot\vec{w}}^{\|\vec{w}\|^2}}{\|\vec{w}\|} + w_0 \text{, or}$$

$$y(\vec{x}) = r\,\|\vec{w}\| \Rightarrow r = \frac{y(\vec{x})}{\|\vec{w}\|}.$$

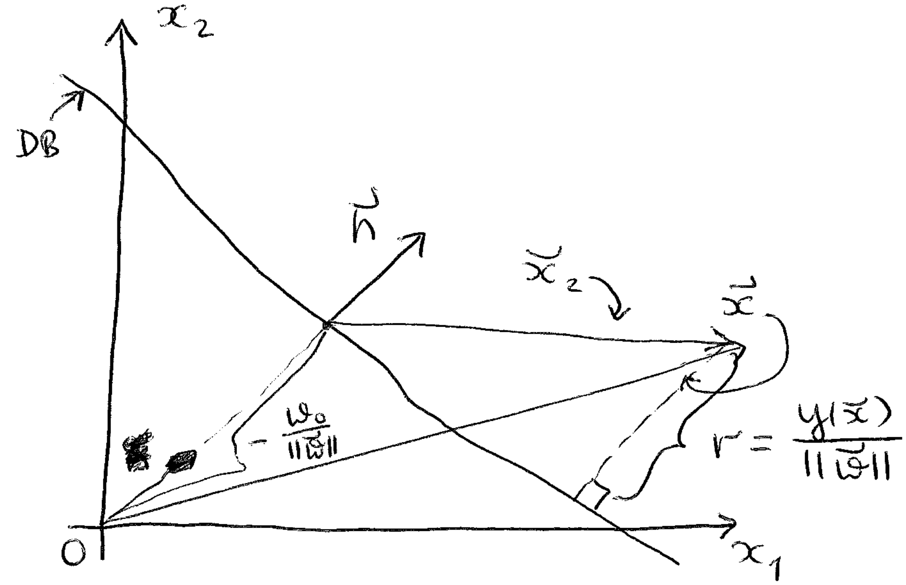
perpendicular distance from $\vec{x}$ to DB

Here, we used

$$\vec{w}^T\cdot\vec{x}_{\parallel DB} + w_0 = 0$$

Indeed,

$\vec{n} \perp DB$



Then

$$-\vec{n}\,\frac{w_0}{\|\vec{w}\|} + \underbrace{\vec{x}_2}_{=\,\vec{x}_{2,\parallel DB} + \underbrace{\vec{x}_{2,\perp DB}}_{=\,r\vec{n}}} = \vec{x}$$

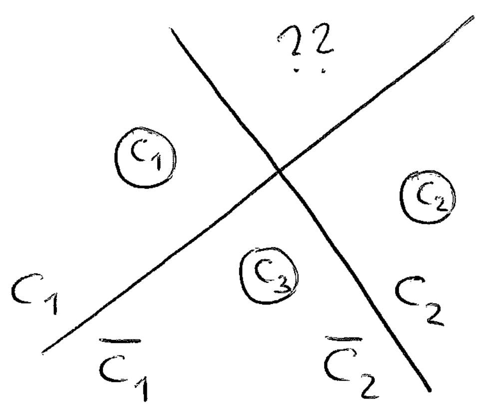
$$\vec{w}^T\cdot\vec{x} = -\underbrace{\frac{\vec{w}^T\cdot\vec{w}}{\|\vec{w}\|^2}}_{=1} w_0 + \underbrace{\vec{w}^T\cdot\vec{x}_{2,\parallel DB}}_{=0} + r\underbrace{\frac{\vec{w}^T\cdot\vec{w}}{\|\vec{w}\|}}_{\|\vec{w}\|}, \text{ or}$$

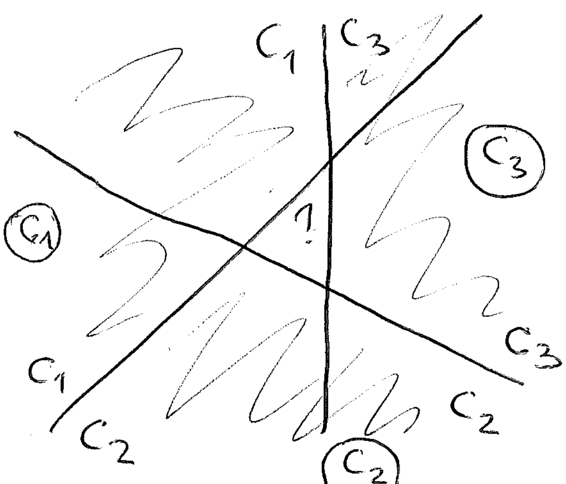
$$\underbrace{\vec{w}^T\cdot\vec{x} + w_0}_{y(\vec{x})} = r\|\vec{w}\|, \text{ as before.}$$

(2.) Multiple classes

Consider $K>2$ classes.

Difficulties in generalizing from the $K=2$ case:





majority vote among discriminant f's

Rather, consider a single K-class discriminant:

$$y_k(\vec{x}) = \vec{w}_k^T\cdot\vec{x} + w_{k,0}$$

Assign a point to C_k if $y_k(\vec{x}) > y_j(\vec{x})$, $\forall j\neq k$.

DBs are then given by $y_k(\vec{x}) = y_j(\vec{x})$, s.t.

$$(\vec{w}_k - \vec{w}_j)^T \cdot \vec{x} + (w_{k,0} - w_{j,0}) = 0$$

Now, consider $\vec{x}_A, \vec{x}_B \in C_k$

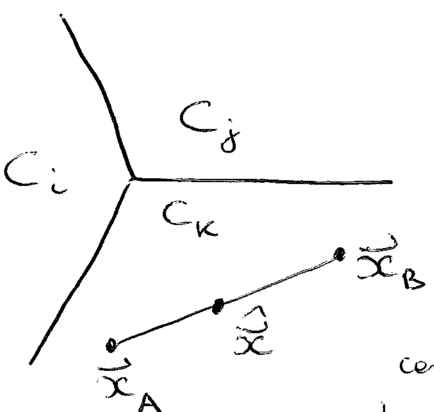
$$\underbrace{\hat{\vec{x}} = \lambda \vec{x}_A + (1-\lambda)\vec{x}_B}_{\text{line connecting } \vec{x}_A \,\&\, \vec{x}_B} \qquad 0 \leq \lambda \leq 1$$

$\Downarrow$ linearity of discriminant f's

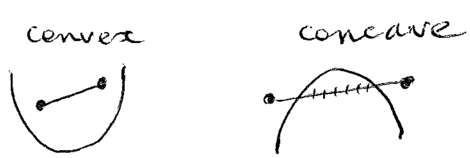
$$y_k(\hat{\vec{x}}) = \lambda \underbrace{y_k(\vec{x}_A)}_{> y_j(\vec{x}_A),\ \forall j \neq k} + (1-\lambda) \underbrace{y_k(\vec{x}_B)}_{> y_j(\vec{x}_B),\ \forall j \neq k}$$

$$y_k(\hat{\vec{x}}) > y_j(\hat{\vec{x}}), \quad \forall j \neq k$$

So, $\hat{\vec{x}} \in C_k$ as well.



Since $\vec{x}_A, \vec{x}_B$ are arbitrary, C_k is singly connected & convex, $\forall k$.



Least squares for classification

Consider a problem with K classes, s.t. $\vec{t}$ are K-dim unit vectors.

$$y_k(\vec{x}) = \vec{w}_k^T \vec{x} + w_{k,0} \Rightarrow \vec{y}(\vec{x}) = \widetilde{W}^T \cdot \tilde{\vec{x}}$$

$$\widetilde{W} = \overbrace{\begin{pmatrix} & w_{k,0} & \\ & w_{k,1} & \\ & \vdots & \\ & \underbrace{w_{k,D}}_{k\text{th column}} & \end{pmatrix}}^{K\ (\#\text{ classes})} \left.\vphantom{\begin{pmatrix}1\\1\\1\\1\end{pmatrix}}\right\} D+1 \ (\#\text{ prms}) \text{ or } \#\text{ entries in } \tilde{\vec{x}}$$

$$\tilde{\vec{x}} = (\underset{x_0}{1}, \vec{x})^T$$

Training set: $\{\vec{x}_n, \vec{t}_n\}$ $n = 1, \dots, N$

Define
$$T = \overbrace{\begin{pmatrix} t_{n,0} & t_{n,1} & \dots & t_{n,K-1} \end{pmatrix}}^{K} \Big\} N$$

$\vec{t}_n$ is the nth row of T

$$\widetilde{X} = \underbrace{\begin{pmatrix} \overset{1}{x_{n,0}} & \dots & x_{n,D} \end{pmatrix}}_{D+1} \Big\} N$$

$\tilde{\vec{x}}_n$ is the nth row of $\widetilde{X}$

$\widetilde{X}\widetilde{W}$ is an $\underset{\#\text{rows}}{N} \times \underset{\#\text{columns}}{K}$ matrix like T

Then $$E(\widetilde{W}) = \frac{1}{2} \mathrm{Tr}\left\{ (\widetilde{X}\widetilde{W} - T)^T (\widetilde{X}\widetilde{W} - T) \right\}.$$

Indeed,

$$E(\widetilde{W}) = \frac{1}{2} \sum_{k=1}^{K} \sum_{n=1}^{N} \Big(\underbrace{(\widetilde{X}\widetilde{W})_{nk}}_{\sum_{j=0}^{D} \widetilde{X}_{nj} \widetilde{W}_{jk} = \sum_{j=0}^{D} \widetilde{x}_{n,j} w_{k,j} = w_{k,0} + \widetilde{w}_k^T \cdot \widetilde{x}} - t_{nk} \Big)^2 \ominus$$

$$\ominus \frac{1}{2} \sum_{k=1}^{K} \sum_{n=1}^{N} \Big(\sum_{j=0}^{D} w_{k,j} \widetilde{x}_{n,j} - t_{nk} \Big)^2 .$$

Then

$$\frac{\partial E}{\partial w_{k,j}} = \frac{1}{2} 2 \sum_{n} \Big[\sum_{j'} w_{k,j'} \widetilde{x}_{n,j'} - t_{nk} \Big] \widetilde{x}_{n,j} = 0, \quad \forall k,j$$

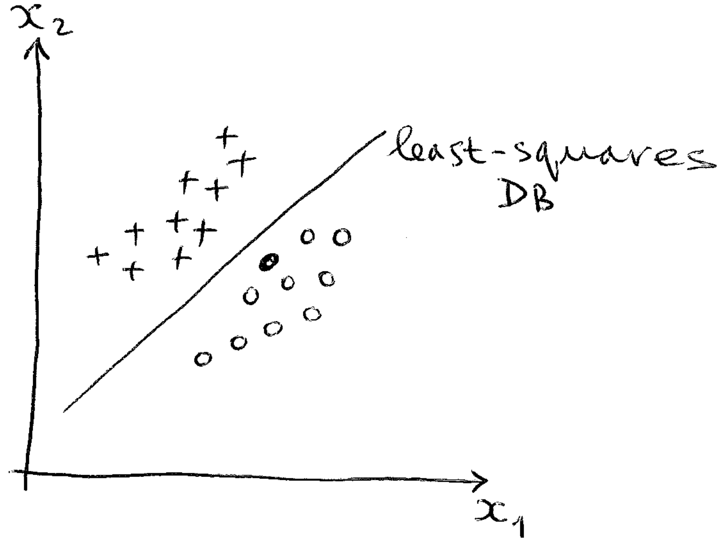
$$\underbrace{\underbrace{(\widetilde{X}^T \widetilde{X})}_{(D+1)\times(D+1)} \underbrace{\widetilde{W}}_{(D+1)\times K}}_{(D+1)\times K} = \underbrace{\widetilde{X}^T T}_{(D+1)\times K}$$

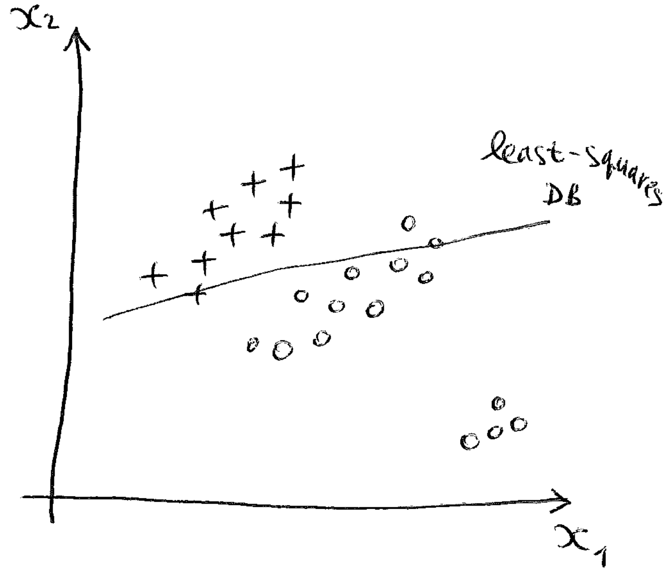
Finally, $$\widetilde{W} = \underbrace{(\widetilde{X}^T \widetilde{X})^{-1} \widetilde{X}^T}_{\text{"} \widetilde{X}^+ \text{ pseudo-inverse of } X} T$$

$$\widetilde{y}(\widetilde{x}) = \widetilde{W}^T \cdot \widetilde{x} = T^T (\widetilde{X}^+)^T \widetilde{x} .$$

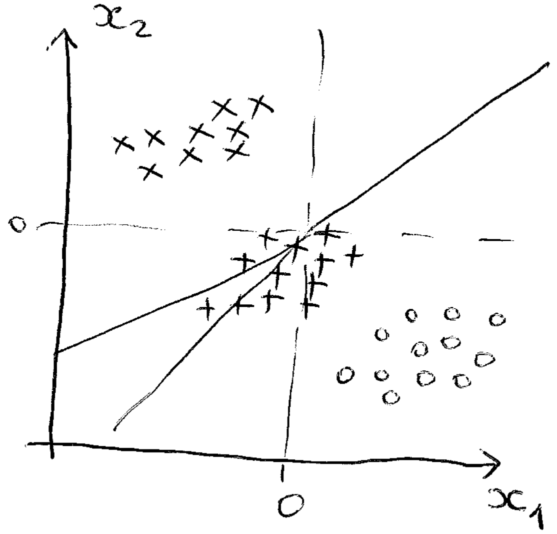
This is a closed-form solution which however is sensitive to outliers.

Ex.





and may even fail <u>completely</u>:



Indeed, least-squares assumes gaussian distr'n of $\vec{t}$'s, and binary target vectors often have non-gaussian distributions.